# Evolutionary Analysis of Scientific Co-authorship Networks in Computer Science: Postprint

**Authors:** Zou Dingjie

**Date:** 2023-10-08T00:00:00+00:00

## Abstract

[Purpose/Significance] To reveal the developmental patterns and characteristics of scientific research collaboration in computer science. [Method/Process] Based on approximately 3 million papers published between 1975 and 2014 in the field of computer science, a co-authorship network is constructed, its evolutionary characteristics are analyzed on an annual basis, and a comparative analysis is conducted between journal paper co-authorship networks and conference paper co-authorship networks. [Results/Conclusion] Research collaboration has become an inevitable trend and widespread phenomenon in computer science, with over 80% of researchers having two or more collaborators. The development of computer science can be divided into two stages—slow development and rapid development—demarcated by the year 2000. Conferences, as a more efficient medium for scientific research information exchange, better facilitate scientific collaboration in the computer science field.

## Full Text

## Analysis on the Evolution of Author Cooperative Network in Computer Science

**Zou Dingjie**
Department of Military Information Management, Shanghai Branch of Nanjing Institute of Politics, Shanghai 200433

### Abstract

[**Purpose/Significance**] This study aims to reveal the developmental patterns and characteristics of scientific collaboration in the field of computer science. [**Method/Process**] We constructed a co-authorship network based on approximately 3 million papers published in computer science from 1975 to 2014, analyzing the evolutionary characteristics of this network on an annual basis. We also conducted a comparative analysis between journal paper co-authorship networks and conference paper co-authorship networks. [**Result/Conclusion**] Research collaboration has become an inevitable trend and universal phenomenon in computer science, with over 80% of researchers having two or more collaborators. The development of computer science can be divided into two stages, with the year 2000 serving as the boundary: a slow development phase and a rapid development phase. As a more efficient medium for scientific information exchange, conferences better facilitate scientific collaboration in computer science research.

**Keywords:** computer science; co-authorship network; evolution analysis

---

With the further development of scientific research, collaboration has become a universal phenomenon across scientific disciplines. Research papers represent the primary manifestation of scientific achievements, and co-authorship patterns can reflect the status of scientific cooperation to a considerable extent. Analyzing co-authorship patterns helps understand the current state of research collaboration, identify its patterns and characteristics, and enables research administrators to strengthen management while inspiring researchers to engage in more effective cooperation.

Foreign scientometrics experts such as Price and D. Beaver pioneered investigations into scientific collaboration and co-authorship patterns. Price argued that most productive authors increase their output due to collective leadership that enables them to accomplish more than they could working alone. Domestic bibliometrics experts Wang Chongde and Wang Bing applied statistical analysis methods to study metrics such as co-authorship rates and degrees, quantitatively examining co-authorship phenomena. Subsequently, researchers including M.E. Newman and A.L. Barabási introduced network approaches to study co-authorship relationships, with social network analysis gaining increasing attention in this domain. Wang Fusheng and Yang Hongyong constructed a co-authorship network based on authors publishing in *Journal of the China Society for Scientific and Technical Information* from 2001-2006, exploring its small-world and scale-free properties. Li Liang and Zhu Qinghua built a co-authorship network from 880 papers published in the same journal between 1998-2005, conducting centrality analysis, cohesive subgroup analysis, and core-periphery structure analysis using social network methods. Later, Liu Bei, Yuan Yi, and colleagues established a collaboration network from 9,806 papers on "information science" retrieved from CNKI between 2000-2006, analyzing characteristics such as collaboration intensity.

However, existing research has primarily focused on library and information science, with small data scales and narrow time spans, typically employing static rather than dynamic analysis. This study constructs co-authorship networks from approximately 3 million papers published in computer science from 1975-2014, analyzing network evolution on an annual basis and comparing differences between journal and conference co-authorship networks to reveal developmental patterns and characteristics in computer science.

## 1 DBLP Database and Data Preprocessing

DBLP is a computer science bibliography database developed and maintained by the University of Trier, Germany. It indexes major international journals and conference proceedings in computer science, providing free literature retrieval services for researchers. With its rapid update cycle and high-quality paper collection, DBLP effectively reflects cutting-edge directions in computer science, enjoys excellent reputation in academia, and has gained high recognition for its authority. As of August 2015, the database had indexed approximately 3.6 million publications by over 1.4 million authors, including about 1.2 million journal articles (46%) and 1.6 million conference papers (53%). This study extracted journal and conference papers published between 1975-2014 from this dataset.

DBLP provides data in XML format, with each record containing fields such as paper title, authors, publication venue, and date. Journal articles are marked with `<article>` nodes, including attributes for creation date (mdate) and unique identifier (key), along with child nodes for authors, title, journal name, and publication year. Conference papers are marked with `<inproceedings>` nodes with similar attributes and child nodes. Due to various reasons, some records lack complete fields. We selected only records with all seven required elements: creation date, unique identifier, authors, title, journal/proceedings name, and publication year, deleting incomplete records. This yielded 1,231,308 journal papers and 1,607,382 conference papers published between 1975-2014. We used Java with the SAX large-scale XML processing package to handle all documents.

[Figure 1: see original paper]

## 2 Co-authorship Network Construction

This study examines the presence or absence of collaboration among researchers without considering collaboration strength, thus constructing an undirected, unweighted co-authorship network. Authors are identified by name and serve as network nodes; an undirected edge is established between any two authors who have co-authored a paper. The network construction process proceeds as follows: (1) Build an "author-coauthor" association table based on paper data; (2) Statistically analyze the current network scale from the association table; (3) Calculate node degrees and degree distributions from the association table; (4) Analyze network connected components using breadth-first search algorithms and calculate the proportion of nodes in the largest component relative to the entire network.

Initially, we constructed a co-authorship network based on papers published in

1975, then expanded to papers from 1975-1976, and continued this process iteratively to ultimately build networks covering papers from 1975-2014, analyzing evolutionary characteristics of network scale and degree distribution over these 40 years.

## 3 Co-authorship Network Evolution Analysis

### 3.1 Overall Network Properties

Table 1 shows basic properties of the journal and conference co-authorship networks constructed from the complete 1975-2014 dataset. The journal co-authorship network includes approximately 930,000 authors who published 1.23 million papers, averaging 3.55 papers per author. The conference co-authorship network includes approximately 1.07 million authors who published 1.6 million papers, averaging 4.52 papers per author. These figures indicate that computer science researchers prefer publishing in conference proceedings to facilitate collaboration and exchange, as conferences provide face-to-face interaction opportunities that benefit scientific communication. The conference network's average of 7.73 coauthors per author exceeds the journal network's 6.90, suggesting researchers are more inclined to collaborate when publishing conference papers—a phenomenon related to the higher timeliness and novelty requirements of conference submissions, which demand more extensive collaboration to complete quality work within shorter timeframes.

From a connectivity perspective, the largest component (connected region) in both networks accounts for over 80% of total nodes, while the second-largest component proportion is extremely low. This indicates that except for a few isolated nodes, the vast majority of authors belong to a single connected region, with increasingly close ties among researchers. The conference co-authorship network's largest component is larger than that of the journal network, indicating better co-authorship connectivity in conference papers.

### 3.2 Network Scale Evolution Analysis

Network scale represents the number of researchers participating in computer science during a given period, reflecting the field's level of attention. Figure 2 shows annual new node additions to both networks, with the x-axis representing years and the y-axis representing new entrants. The figure reveals that newly added researchers have been increasing annually, with the trend accelerating under the wave of informatization. Computer science development can be divided into two stages: 1975-2000, characterized by slow growth with comparable growth rates between journal and conference networks; and 2000-2014, marked by explosive growth, particularly in conference papers. This growth primarily stems from the rise of internet and information technologies around 2000, which attracted numerous researchers to computer science and stimulated exponential growth in new participants. Additionally, conferences' advantages in facilitating face-to-face communication and greater timeliness have consistently attracted more researchers than journals.

[Figure 2: see original paper]

### 3.3 Degree Distribution Evolution Analysis

Node degree refers to the number of adjacent nodes. Degree distribution represents the ratio of nodes with degree $n$ to total network nodes. In co-authorship networks, a node represents a researcher, node degree represents their number of collaborators, and average degree represents the mean number of collaborators per researcher. Degree distribution provides more detailed information than average degree about collaboration patterns in the field.

Figure 3 shows the evolution of average degree in both networks, demonstrating a linear increase over time. This indicates that collaboration has become the dominant trend in computer science, making solo work or limited collaboration inefficient for high-impact research. Conference network researchers consistently show higher collaboration numbers than their journal counterparts, primarily due to conferences' shorter production cycles requiring more researchers to complete quality work within limited timeframes.

Degree distribution offers more granular insight into researcher collaboration patterns. Overall, over 80% of researchers have six or fewer collaborators, so we focus on nodes with degrees 0-6. Figures 4 and 5 present degree distribution evolution, with x-axes representing years and y-axes showing the proportion of nodes with degree $n$. The trends show that proportions of nodes with degree 0 or 1 decline annually, while degree-2 nodes initially increase then decline after 2000. Nodes with degrees 3-6 show clear upward trends. These patterns indicate that researchers working alone or with only one collaborator are increasingly rare, while those with two or more collaborators are becoming the norm. Researchers with 2-4 collaborators now exceed 50% and continue rising, establishing this as the mainstream pattern. Those with 5-6 collaborators are also increasing, suggesting large-scale collaboration will become a future trend.

[Figure 3: see original paper]
[Figure 4: see original paper]
[Figure 5: see original paper]

### 3.4 Largest Component Evolution Analysis

In early-stage social networks, sparse connections typically produce isolated nodes or small isolated regions. Over time, existing nodes strengthen connections and new nodes link to existing ones, gradually forming a giant connected component alongside several very small isolated regions. These regions are called components, each representing a community in the network. This section examines component evolution in computer science co-authorship networks.

Computer science co-authorship networks tend to evolve into a single integrated network. Initially composed of many small components, the network becomes increasingly cohesive over time, ultimately forming one giant component and several minimal ones. The journal co-authorship network contained 1,718 components in 1975, with the largest component comprising nearly 0.48% of nodes and the second-largest 0.47%, while approximately 200 components each exceeded 0.1%—demonstrating a fragmented structure of numerous small, isolated components. By 2014, the journal network's largest component had grown to 81% of nodes while the second-largest component shrank to $2.5\times10^{-5}$, indicating that the vast majority of researchers had merged into a giant component with only scattered individuals remaining isolated. This evolution demonstrates increasingly close collaboration and healthy development of scientific cooperation.

Figure 6 shows the temporal evolution of largest component proportions in both networks. Starting from approximately 3%, the largest component gradually grew to over 80%. In later development stages, growth slows around 90% because approximately 10% of nodes remain isolated. This slowdown stems from two factors: (1) single-paper authors who emerge annually and permanently remain as network isolates, making theoretically complete connectivity impossible; and (2) some new nodes cannot immediately connect to the largest component upon entry, forming temporarily isolated regions.

[Figure 6: see original paper]

## 4 Conclusions and Implications

Based on 3 million papers from the DBLP dataset (1975-2014), this study constructed scientific collaboration networks and analyzed computer science co-authorship evolution, revealing patterns and characteristics through four perspectives: basic network properties, network scale, degree distribution, and largest component evolution.

Co-authorship in computer science exhibits three key features: (1) Collaboration has become a universal phenomenon and inevitable trend, representing the primary mode of scientific research. By 2014, the proportion of researchers publishing independently had declined to 3% in the journal network and 2% in the conference network, while those with 2-4 collaborators exceeded 50% and continues rising. The evolution toward a giant connected component also reflects increasingly close researcher relationships. (2) Computer science development divides into two stages: slow growth (1975-2000) and explosive growth (2000-2014). (3) Conference papers better promote researcher collaboration than journal papers. Conference network growth rates exceed journal network rates, with conferences attracting more participants. Post-2000, the conference network's largest component grew faster than the journal network's, and its consistently higher average degree indicates closer collaboration relationships.

Three factors may explain these characteristics. First, the need to jointly solve complex problems drives collaboration. Early-stage research addresses fundamental questions requiring fewer personnel, while later-stage complex problems demand more extensive cooperation. Second, market demands for computers and internet technologies since 2000 have stimulated field development, attracting numerous researchers. Third, conferences' higher timeliness requirements objectively necessitate stronger collaboration to complete quality work within shorter cycles, while face-to-face communication opportunities at conferences may further attract researcher participation.

Large-scale co-authorship network analysis based on massive datasets can more accurately and comprehensively reveal disciplinary collaboration patterns, but faces significant data processing challenges that make traditional personal computers inadequate for analyzing giant networks. Future work will explore efficient methods for massive data processing and giant network analysis.

## References

[1] Price D. *Little Science, Big Science* [M]. Translated by Song Jiangeng, Dai Zhenfei. Beijing: World Science Press, 1982.
[2] Beaver D, Rosen R. Studies in scientific collaboration: part I. the professional origins of scientific co-authorship[J]. *Scientometrics*, 1978, 1(1): 65-84.
[3] Wang Chongde. Research on scientific paper authors[J]. *Journal of the China Society for Scientific and Technical Information*, 1982, 1(2): 220-225.
[4] Wang Bing. Research on co-authorship phenomena in Chinese natural science journals[J]. *Journal of Information Science*, 1990, 11(5): 335-339.
[5] Newman M E. The structure of scientific collaboration networks[J]. *Working papers*, 2000, 98(2): 404-409.
[6] Barabási A L, Jeong H, Néda Z, et al. Evolution of the social network of scientific collaborations[J]. *Physica A: Statistical Mechanics and its Applications*, 2002, 311(3-4): 590-614.
[7] Wang Fusheng, Yang Hongyong. Analysis of author collaboration network in *Journal of the China Society for Scientific and Technical Information*[J]. *Journal of the China Society for Scientific and Technical Information*, 2007, 26(5): 659-663.
[8] Li Liang, Zhu Qinghua. Empirical study of social network analysis in co-authorship analysis[J]. *Information Science*, 2008(4): 549-555.
[9] Zhu Qinghua, Li Liang. Social network analysis and its application in information science[J]. *Information Studies: Theory & Application*, 2008, 31(2): 179-183.
[10] Liu Bei, Yuan Yi, Boutin E. Application of social network analysis in research paper collaboration networks[J]. *Journal of the China Society for Scientific and Technical Information*, 2008, 27(3): 407-417.
[11] [EB/OL]. [2015-11-08]. http://dblp.uni-trier.de/db/.
[12] Dou Binglin, Li Susong, Zhang Shiyong. Structural analysis of social networks[J]. *Chinese Journal of Computers*, 2012, 35(4): 741-753.

*Note: Figure translations are in progress. See original paper for figures.*